A note on ADMM

Tri Nguyen
nguyetr9@oregonstate.edu

March 18, 2022

I encounter ADMM several time, did go through the derivation 2 times at least, still I cannot write it down from scratch when a friend of mine asked me about it. This note is to summary my understand about it.

1 Dual Ascend

Firstly, take a look at a common optimization problem with linear equality constraint

$$\underset{x}{\text{minimize}} \quad f(x) \tag{1a}$$

$$\text{subject to} \quad Ax = b, \tag{1b}$$

The first method comes to mind is Lagrangian multiplier. The Lagrangian is

$$L(x, y) = f(x) + y^\top (Ax - b).$$

Optimizing on $L(x, y)$ is easier since it is a unconstrained optimisation problem.

$$\underset{x,y}{\text{minimize}} \; L(x, y) \tag{2}$$

Solutions (x^*, y^*) of this problem would potentially provide us solution of the original problem.

Block 1.1. *If x^* is a solution of Problem* (1)*, then it would be a part of a stationary point of problem* (2)*, i.e, (x^*, y^*) is a stationary point of problem* (2) *for some y^*.*

Proof. It might be too much but I will invoke KKT conditions. KKT conditions provides a set necessary conditions for a stationary point of a constrained problem. One of the condition is

$$\frac{\partial\, L(x^*, y)}{\partial\, x} = 0$$

Hence, any solution x^* of problem 1 would be stationary point of problem (2).

□

Block 1.2 (KKT conditions. (no proof))**.** *Necessary condition and sufficient condition:*

- *For a constrained optimisation problem with zero duality gap, if x^* and (λ^*, ν^*) are primal and dual optimal then x^*, λ^*, ν^* are satisfied KKT conditions.*
- *For a constrained convex optimisation problem, if x^*, λ^*, ν^* are satisfied KKT conditions, then x^* and (λ^*, ν^*) are primal and dual optimal with zero duality gap.*

Block 1.3 (Strong duality)**.** *These conditions which guarantees strong duality are called constraint qualifications.*

- *If the problem is convex, Slater's condition is a constraint qualification. It requires existence of a strictly feasible point.*

- *Other constraint qualifications.*

So one can work on the Lagrangian problem then pick out the best solution for the original problem. That fact that solving Problem (2) is much easier makes this method a good-to-go method to deal with this constrained problem.

Block 1.4 (Lagrangian multiplier)**.**

2 Derivation

It's not really a derivation but rather a step by step how to get to the procedure above.

1. The augmented Lagrangian function:

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

2. The dual function:

$$g(y) := \inf_{x,z} L_\rho(x, z, y)$$

3. Dual problem:

$$\underset{y}{\text{maximize}} \quad g(y)$$

using gradient ascent, where the gradient respect to y is

$$\nabla g(y) = Ax + Bz - c \quad \text{(prove this)}$$

So 3 steps are

$$\begin{aligned}
x^{k+1} &\leftarrow \arg\min_x \quad L_\rho(x, z^k, y^k) \\
z^{k+1} &\leftarrow \arg\min_z \quad L_\rho(x^{k+1}, z, y^k) \\
y^{k+1} &\leftarrow y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)
\end{aligned}$$

which are equivalent to

$$\begin{aligned}
x^{k+1} &\leftarrow \arg\min_x \quad \left(f(x) + \frac{\rho}{2}\left\|Ax + Bz^k - c + u^k\right\|\right) \\
z^{k+1} &\leftarrow \arg\min_z \quad \left(g(z) + \frac{\rho}{2}\left\|Ax^{k+1} + Bz - c + u^k\right\|\right) \\
u^{k+1} &\leftarrow u^k + Ax^{k+1} + Bz^{k+1} - c
\end{aligned}$$

where $u = \frac{1}{\rho}y$.
What if we define residual as $-Ax - Bz + c$?

3 Examples

3.1 Problem 1

$$\begin{aligned}
\underset{\mathbf{A}}{\text{minimize}} \quad & \left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C} \odot \mathbf{B})\text{diag}\left(\boldsymbol{\lambda}\right)\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\
\text{subject to} \quad & \mathbf{A} \geq 0, \mathbf{1}^{\mathrm{T}}\mathbf{A} = \mathbf{1}^{\mathrm{T}}
\end{aligned}$$

We can use PGD to solve this, but ADMM is more flexible, and we are not sure which one is faster.

- Step 1: Translate problem to ADMM form

$$\begin{aligned} \underset{\mathbf{A},\widetilde{\mathbf{A}}}{\text{minimize}} \quad & \frac{1}{2}\left\|\underline{\mathbf{X}} - (\mathbf{C}\odot\mathbf{B})\text{diag}\left(\boldsymbol{\lambda}\right)\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + I(\widetilde{\mathbf{A}}\in\Delta) \\ \text{subject to} \quad & \mathbf{A} = \widetilde{\mathbf{A}} \end{aligned}$$

- Step 2: Plug in ADMM updating rule

$$\begin{aligned} \mathbf{A}^{k+1} &\leftarrow \underset{\mathbf{A}}{\text{minimize}} \quad \left(\frac{1}{2}\left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C}\odot\mathbf{B})\text{diag}\left(\boldsymbol{\lambda}\right)\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \frac{\rho}{2}\left\|\mathbf{A} - \widetilde{\mathbf{A}}^{k} + \mathbf{U}^{k}\right\|_{\mathrm{F}}^{2}\right) \\ \widetilde{\mathbf{A}}^{k+1} &\leftarrow \underset{\widetilde{\mathbf{A}}}{\text{minimize}} \quad I(\widetilde{\mathbf{A}}\in\Delta) + \frac{\rho}{2}\left\|\mathbf{A}^{k+1} - \widetilde{\mathbf{A}} + \mathbf{U}^{k}\right\|_{\mathrm{F}}^{2} \\ &= \underset{\widetilde{\mathbf{A}}}{\text{minimize}} \quad I(\widetilde{\mathbf{A}}\in\Delta) + \frac{\rho}{2}\left\|\widetilde{\mathbf{A}} - \mathbf{A}^{k+1} - \mathbf{U}^{k}\right\|_{\mathrm{F}}^{2} \\ &= \mathcal{P}_{\Delta}(\mathbf{A}^{k+1} + \mathbf{U}^{k}) \\ \mathbf{U}^{k+1} &\leftarrow \mathbf{U}^{k} + \mathbf{A}^{k+1} - \widetilde{\mathbf{A}}^{k+1} \end{aligned}$$

  Solve for $\mathbf{A}^{\mathrm{T}}$:

$$\begin{aligned} & \nabla = \mathbf{0} \\ \Leftrightarrow & \text{diag}\left(\boldsymbol{\Lambda}\right)(\mathbf{C}\odot\mathbf{B})^{\mathrm{T}}\left((\mathbf{C}\odot\mathbf{B})\text{diag}\left(\boldsymbol{\Lambda}\right)\mathbf{A}^{\mathrm{T}} - \underline{\mathbf{X}}^{(1)}\right) + \rho(\mathbf{A}^{\mathrm{T}} - \left(\widetilde{\mathbf{A}}^{k}\right)^{\mathrm{T}} + \left(\mathbf{U}^{k}\right)^{\mathrm{T}}) = \mathbf{0} \\ \Leftrightarrow & \left(\text{diag}\left(\boldsymbol{\Lambda}\right)(\mathbf{C}^{\mathrm{T}}\mathbf{C} * \mathbf{B}^{\mathrm{T}}\mathbf{B})\text{diag}\left(\boldsymbol{\Lambda}\right) + \rho\mathbf{I}_{\text{size}(\mathbf{A}^{\mathrm{T}},\,1)}\right)\mathbf{A}^{\mathrm{T}} = \text{diag}\left(\boldsymbol{\Lambda}\right)(\mathbf{C}\odot\mathbf{B})^{\mathrm{T}}\underline{\mathbf{X}}^{(1)} + \rho\left(\left(\widetilde{\mathbf{A}}^{k}\right)^{\mathrm{T}} - \left(\mathbf{U}^{k}\right)^{\mathrm{T}}\right) \\ & \mathbf{G}\mathbf{A}^{\mathrm{T}} = \mathbf{H} \end{aligned}$$

3.2 Example 2

$$\begin{aligned} \underset{\mathbf{A}}{\text{minimize}} \quad & \left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C}\odot\mathbf{B})\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} \\ \text{subject to} \quad & \mathbf{A}\in\mathcal{C} \end{aligned}$$

Step 1: translate to ADMM form. There are several ways to do that, and they are slightly different. Let's stick with the following recipe. The little transpose notation is very easily confusing.

$$\begin{aligned} \underset{\mathbf{A},\widetilde{\mathbf{A}}}{\text{minimize}} \quad & \left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C}\odot\mathbf{B})\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^{2} + I(\mathbf{A}\in\mathcal{C}) \\ \text{subject to} \quad & \mathbf{A} = \widetilde{\mathbf{A}}^{\mathrm{T}} \end{aligned}$$

Step 2:

- Update $\widetilde{\mathbf{A}}$,

$$\begin{aligned} \widetilde{\mathbf{A}} &\leftarrow \frac{1}{2}\left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C}\odot\mathbf{B})\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^{2} + \frac{\rho}{2}\left\|\mathbf{A} - \widetilde{\mathbf{A}}^{\mathrm{T}} + \mathbf{U}\right\| \\ &\leftarrow \frac{1}{2}\left\|\underline{\mathbf{X}}^{(1)} - (\mathbf{C}\odot\mathbf{B})\widetilde{\mathbf{A}}\right\|_{\mathrm{F}}^{2} + \frac{\rho}{2}\left\|\widetilde{\mathbf{A}} - \mathbf{A}^{\mathrm{T}} - \mathbf{U}^{\mathrm{T}}\right\| \\ & \nabla_{\widetilde{\mathbf{A}}} = 0 \\ \Leftrightarrow & ((\mathbf{C}^{\mathrm{T}}\mathbf{C} * \mathbf{B}^{\mathrm{T}}\mathbf{B}) + \rho\mathbf{I})\widetilde{\mathbf{A}} = (\mathbf{C}\odot\mathbf{B})^{\mathrm{T}}\underline{\mathbf{X}}^{(1)} + \rho(\mathbf{A} + \mathbf{U})^{\mathrm{T}} \\ \Leftrightarrow & \widetilde{\mathbf{A}} = \left[(\mathbf{C}^{\mathrm{T}}\mathbf{C} * \mathbf{B}^{\mathrm{T}}\mathbf{B}) + \rho\mathbf{I}_{F}\right]^{-1}\left[(\mathbf{C}\odot\mathbf{B})^{\mathrm{T}}\underline{\mathbf{X}}^{(1)} + \rho(\mathbf{A} + \mathbf{U})^{\mathrm{T}}\right] \end{aligned}$$

- Update $\mathbf{A}$ usually with a proximal operator,

$$\mathbf{A} \leftarrow \arg\min_{\mathbf{A}} \; I(\mathbf{A}\in\mathcal{C}) + \frac{\rho}{2}\left\|\mathbf{A} - \widetilde{\mathbf{A}}^{\mathrm{T}} + \mathbf{U}\right\|_{\mathrm{F}}^{2}$$

- Update dual variable $\mathbf{U}$,

$$\mathbf{U} \leftarrow \mathbf{U} + \mathbf{A} - \widetilde{\mathbf{A}}^{\mathrm{T}}$$

Does the order of update $\mathbf{A}, \widetilde{\mathbf{A}}$ matter? How's about initilization for each sub-problem? Should we init them randomly, or used value from previous iteration?
Small ρ is important to converge to a smaller objective value.